**Abstract**

Brains enjoy a bodily life. Therefore animals are subjects with a point of view. Yet, coding betrays an anthropomorphic bias: we can, therefore they must. Here I propose a reformulation of Brette's question that emphasizes organismic perception, cautioning for misinterpretations based on external ideal-observer accounts. Theoretical ethology allows computational neuroscience to understand brains from the perspective of their owners.

An apparently innocuous word in Brette's question is a major source of confusion but also contains a great deal of the answer. Is coding a relevant metaphor for "the" brain? Yes and no. It depends on whose brain we are talking about. For the scientist studying the animal, coding is certainly relevant (at least, as the ubiquity of such figure of speech attests in current neuroscience). But, insofar as we are interested in the animal and its brain, the answer is likely no. The mantra "stimulate, record, correlate" misses the point of the organism. It is *for us, by us*. That the experimenter's model can decode the signal does not mean that the brain can or does. The information necessary to make sense of the data in terms of coding is seldom available to the organism, upon which coding is predicated. This creates a can-ought problem: a description of what the neuroscientist can do prescribes what the animal must do. Such implicit tension pervades most of the disagreements that Brette's question shall spur. The problem, I believe, is deeper than coding: There is a conflict of interests between the scientist and the laboratory animal.

Biology is the science of living beings. Organisms are centers of action. As such, perspective matters. To be an organism is to have a point of view. All animals share a common world but not all animals have a world in common. Each living organism has its own *Umwelt* (meaningful environment), which is different than its *Umgebung* (physical surroundings): A tree is a tree, but a tree for an ant has little to do with a tree for a carpenter (Uexküll 1926). What is meaningful for an organism – or even what is possibly apprehensible – need not be meaningful for the scientist studying it, and *vice versa* (a concrete and pervasive example: stimuli are more the experimenter's output than the animal's input). The use of the definite article ("*the* brain") or the indefinite pronoun ("*one* finds") is so delicate in biology. It easily blurs the subject (I? you? the mouse? what mouse?), unbinding grave conflations and misleading thought and interpretation. Eloquently said, "Hedgehogs as such do not cross roads (…). On the contrary, it is man-made roads that cross the hedgehog's millieu" (Canguilhem 2008, p. 22). Rather than being an exception, coding illustrates such misattribution. Paraphrasing, we could say that cat brains as such do not encode stripes, but it is stripes that we decode from the cat's brain. A clash of *Umwelts* (*Umwelten*, in proper German) is going on in our laboratories.

The notion of *Umwelt* has no place in physics; it does not violate physics, but it is not reducible to physics either. Living beings inhabit a world of meaning that includes but exceeds the physical world of masses and forces, and even more so the mathematical world of zeros and ones. The appreciation of the uniqueness of biology discords with a cornerstone of the scientific approach: objectivity. Of course we always observe reality from a viewpoint, explicitly or implicitly chosen. But it is ultimately deemed irrelevant. Objectivity, then, is the pretense of self-exclusion from the phenomenon under study. The observer vanishes in classical physics (also in biology). By means of a representation of things that ultimately does not depend on the reference system, an observer-independent reality is erected. Yet, "[o]n the strength of the immediate testimony of our bodies *we* are able to say what no disembodied onlooker would have a cause for saying: (…) the point of life itself: its being self-centered individuality" (Jonas 2001, p. 79). From subjectivity we have prodigiously built an objectivity that can dispense with the former. However, upon inspection, objectivity becomes a particular kind of intersubjective consensus. This is biology's scotoma: We are subjects whose objects of study are subjects too.

In behavioral neuroscience there is an observer-observed gap. Physiology aspires to study the inner workings (brain) of an organism from the outside (scientist's perspective); ethology strives to understand the outer happenings (behavior) from the inside (animal's perspective). Isn't the neurophysiologist's decentering a covert self-centering? Sticking electrodes is not sufficient to know what it is like to be a rat. But, how to look through the animal's eyes? A cute example is *Turtle Geometry*: it actually matters if a turtle traces a circle by solving the $x^2 + y^2 = r^2$ equation, or by iterating a "run and turn" procedure. Both are mathematically equivalent (from an external ideal observer, perhaps indistinguishable, even irrelevant) but biologically they are not the same. There is much to gain from discovering "the range of complicated things a turtle can do in terms of the simplest things it knows" (diSessa & Abelson 1981, p. 3). What is it to make sense from the animal's perspective when it does not do so the way we do? Such is the paradox: The *Umgebung*, the objective world of scientists, can be part of our human *Umwelt* (we do not feel neutrinos crossing our bodies, but we can detect them in bubble chambers), but it collides with the *Umwelt* of the animal, which is never an *Umgebung*. Neuroscientists yearn for neural codes; the animal has no clue.

Neuro-ethology is actually meta-engineering: our problem is to solve how animals solve their problems – to scientifically empathize with each creature. This entails a revision of Bernard's (1957, p. 103) foundational words: The scientist "no longer hears the cry of animals, he no longer sees the blood that flows, he sees only his idea and perceives only organisms concealing problems which he intends to solve." By reformulating Brette's question, my intention here has been to emphasize that computational neuroscience can benefit from the insights of theoretical ethology to transform its anthropomorphic bias. To crack codes, "it would suffice that we be angels. But to do biology, even with the aid of intelligence, we sometimes need to feel like beasts ourselves" (Canguilhem 2008, p. xx). The question then is not so much whether coding is relevant or wrong, but to what extent it is misleading. We must then ask: Whose brain is the coding metaphor relevant for?

**Acknowledgments.** I thank Ehud Ahissar, Konrad Kording, Spyridon Koutroufinis, Ibrahim Tastekin, Zach Mainen, and specially Asif Ghazanfar for insightful discussions.

# Beyond metaphors and semantics: A framework for causal inference in neuroscience

Roberto A. Gulli

Mortimer B. Zuckerman Mind Brain Behavior Institute, Columbia University, New York, NY 10027; Center for Theoretical Neuroscience, Columbia University,

New York, NY 10027; Department of Neuroscience, Columbia University, New York, NY 10027
r.gulli@columbia.edu
http://robertogulli.com

doi:10.1017/S0140525X19001389, e230

**Abstract**

The long-enduring coding metaphor is deemed problematic because it imbues correlational evidence with causal power. In neuroscience, most research is correlational or conditionally correlational; this research, in aggregate, informs causal inference. Rather than prescribing semantics used in correlational studies, it would be useful for neuroscientists to focus on a constructive syntax to guide principled causal inference.

In his article, Brette argues that the "coding metaphor" in neuroscience is inappropriate and misleading because it leads to false interpretations of causality. Brette states that "by postulating that neural codes are representations, we imply that these codes have a causal impact on the brain" (sect. 4.2, para. 1). However, this is implausible since "[in the] technical sense … the word *code* is used as a synonym for correlate" (sect. 1, para. 4). Restated, the coding metaphor is problematic because it can imply causal function where sufficient evidence to support causal inference does not exist. By relying on this criticism, Brette commits to a broader error: He interprets that isolated correlations, conditional correlations, and statistical inferences between neural activity and function support or refute causal inference. Isolated pieces of correlational or statistical evidence are insufficient to demonstrate a causal relationship between neural activity and functions, perceptions, or behaviors, and should be considered in aggregate to form the basis of causal inference. For this reason, it would be helpful for those seeking to design and interpret experiments to adopt a constructive framework for causal inference in neuroscience.

The correlational nature of individual studies in neuroscience has been explicit since the dawn of electrophysiology, when Caton (1875) stated that "[t]he electric currents of grey matter appear to have a relation to its functions." Contemporary studies of neural activity and function are still strictly correlational, despite advances in recording and analysis methods. Traditional statistical techniques are agnostic to causal relationships between variables and thus cannot determine causality (Pearl et al. 2016). Experimental interventions that support causal inferences between brain (dys-)function and behavior have long been sought (Dodds 1878; Ferrier 1886). However, even studies that use modern versions of these "causal" techniques (optogenetic, chemogenetic, electrical, and pharmacological modulation) provide correlations conditioned on perturbation. Causal inferences on the basis of single experimental results should be tempered because of plausible confounding and off-target effects (Jazayeri & Afraz 2017).

Insights from other fields provide a clear path toward causal inference with individually circumstantial pieces of evidence. The most influential perspective may be that of medical statistician Austin Bradford Hill, who described nine "viewpoints" that guide causal inference in epidemiology when randomized controlled trials are not possible (Hill 1965; see also Phillips & Goodman 2004). Here these viewpoints are adapted to form a Bradford Hill-inspired framework for causal inference in neuroscience, where aggregated observational and interventional studies support causal inference:

1. **Correlational evidence:** Relationships between measurements of neural activity and experimenter-defined responses (whether in downstream neural activity, other physiological or behavioral outcomes). These relationships can be characterized through a variety of forward and backward modeling techniques (see, e.g., Anderson 2019; Baayen et al. 2008; Marinescu et al. 2018; Rougier 2019; Saxena and Cunningham 2019; Song et al. 2013; Wang and Yang 2016).
    i. *Strength:* Does the neural activity explain a reasonable amount of variability in the response?
    ii. *Consistency:* Does the neural activity reliably produce the outcome?
    iii. *Specificity:* Is the observed relationship between neural activity response unique or one of a vast array of potentially confounding correlations?
    iv. *Relationship curve:* Is there a clear geometric relationship between neural activity and the response?
    v. *Temporality:* Does the neural activity consistently precede the response in time?
    vi. *Mechanistic plausibility:* Is there a plausible mechanism whereby neural activity may produce response?

2. **Conditionally correlational evidence:** The effect of direct or indirect modulation of neural activity on experimenter-defined outcomes. Modulation includes loss-of-function and gain-of-function "causal" manipulations that are under control of the experimenter.
    i. *Strength:* Does modulation of neural activity explain a reasonable amount of variability in the response?
    ii. *Consistency:* Does modulation of neural activity reliably produce the predicted outcome?
    iii. *Specificity:* Does modulation of neural activity lead to a prescribed outcome or one of a vast array of potential effects?
    iv. *Relationship curve:* Is there a predictable and replicable geometric relationship between modulation of neural activity and the response?
    v. *Temporality:* Does the predicted effect follow the perturbed neural activity at a reasonable delay?
    vi. *Coherence:* Is the predicted effect of modulation of neural activity coherent with other strong hypotheses?
    vii. *Analogy:* Does a modulation of closely related neural activity patterns produce similar effects?

With this framework in mind, one should reconsider Brette's claims related to neural codes and causal inference. For example, Brette states that "BOLD (blood oxygen level-dependent) signal … encodes visual signals in the same technical sense that the firing of neurons encodes visual signals" (sect. 4.2, para. 2). Functional magnetic resonance imaging and electrophysiology studies are both correlational, but Brette's assertion is deeply flawed in important ways. In fMRI and electrophysiology, fundamentally different biological activity is associated with stimulus or behavioral response of interest (Goense and Logothetis 2008). Thus, the mechanistic plausibility of a link between neural activity and the experimental condition differs. Furthermore, the spatial specificity and temporality of visually evoked activity cannot be similarly addressed across techniques (Sejnowski et al. 2014). These factors are critically important in guiding causal inference, and therefore, each technique uniquely contributes toward causal inference. To suggest that BOLD signals and action potentials encode visual stimuli in the same technical sense is a conspicuous oversimplification. In this example, the

proposed framework for causal inference aids in articulating the relative strengths and weaknesses of different experimental approaches. Furthermore, it provides guidelines for making causal inferences by aggregating individual pieces of evidence that are insufficient in isolation.

Regarding the causal relationship between the physical world and thought, Haugeland (1985, p. 106) stated, "If you take care of the syntax, the semantics will take care of itself." This axiom presents a useful analogy: with a proper framework to describe the syntax (rules and criteria) of causal inference in neuroscience, Brette's claim – the coding metaphor perpetuates inappropriate causal inference – is reduced to an innocuous semantic debate. His further claim that metaphors perpetuate "semantic drift" (sects. 2.1, 2.2, and 3.1) should be addressed not by further semantic prescriptions, but by adhering to reasoned syntax. These semantic debates distract from the ultimate goal of discovering robust, causal relationships between the many levels of organization in the brain and behavior.

**Acknowledgments.** I thank David L. Barack, Matthew L. Leavitt, and Rishi Rajalingham for criticism and comments.

# Codes, communication and cognition

Stevan Harnad

Department of Psychology, Université du Québec à Montréal, Montréal (Québec) H3C 3P8, Canada; Department of Electronics and Computer Science, University of Southampton, Southampton SO17 1TW, UK.
harnad@soton.ac.uk

doi:10.1017/S0140525X19001481, e231

**Abstract**

Brette criticizes the notion of neural coding because it seems to entail that neural signals need to "decoded" by or for some receiver in the head. If that were so, then neural coding would indeed be homuncular (Brette calls it "dualistic"), requiring an entity to decipher the code. But I think Brette's plea to think instead in terms of complex, interactive causal throughput is preaching to the converted. Turing (not Shannon) has already shown the way. In any case, the metaphor of neural coding has little to do with the symbol grounding problem.

Both Shannon's (1948) *information* and Turing's (1936) *computation* are important in cognitive science. Shannon is concerned with the faithfulness of signal transmission in communication, and Turing is concerned with what algorithms can do. Cognitive science is concerned with what organisms (hence their brains) can *do*, and *how*.

Cells (including neurons) transmit signals. This is already true in plants (Baluska & Mancuso 2009) and of course also in machines. And organisms certainly do things. Which of the things organisms do are "cognitive" and which are "vegetative" is mostly just a definitional matter, but it is probably overstretching the notion to say that paramecia or hearts are "cognizing." The examples are nevertheless instructive for cognitive science, because paramecia, hearts, and organisms with brains are all systems that can *do* things. So are computers and robots, for that matter. Hence finding a causal explanation of how one of them does what it does may provide useful lessons for explaining the others.

Let's start with the heart, an example used by Brette. What does the heart do? It pumps blood. No metaphors. The heart literally pumps blood, and cardiac science has successfully reverse-engineered the heart (to a close approximation). We know how the heart does it – and part of the proof that we know how is that we can apply and test our hypotheses about how the heart pumps blood by building a synthetic model of a heart, plugging it into the heart's inputs and outputs, and testing whether it can pump blood. If it can, the artificial heart passes the "Turing Test" for cardiac function.

So what does the (human) brain (and body) pump? Human behavior. Or, rather, human behavioral *capacity*. What people *can do*. Let's forget about what portion of that capacity counts as cognitive and what proportion is just vegetative (like cardiac function): It all consists of the capacity of a (living) system to do certain things. Now the challenge is to explain how.

Turing (1950) provided the ground rules: You have an explanation if you can design a system that can do everything a human being can do, indistinguishably – *to* a human – *from* a human. If your interest is just in "cognitive" capacities, then just generate those, ignoring the vegetative capacities (or at least those that are not essential for generating the cognitive capacities). Cognition, like Justice Potter Stewart's pornography, may be hard to define, but we know it when we see it. And the capacity to interact with the dynamic world of objects and events and their properties (including words describing those objects, events, and properties) indistinguishably from the way humans do is surely cognitive, if anything is.

There is one more thing: Humans don't just do: They also feel. It *feels like something*, to a human, to be seeing and doing what humans can see and do. But the capacity to feel eludes Turing's program for cognitive science. It's something our brains pump invisibly. Turing (1950) accordingly brackets it. But it keeps making disruptive peekaboo appearances in our attempts to reverse-engineer cognition, as we shall see.

One of the main hypotheses about how the brain pumps cognitive capacity is via computation, Turing computation. Computation is the manipulation of "symbols" (arbitrary formal objects) on the basis of rules operating only on the symbols' shapes ("syntax"), not their meanings ("semantics"), to generate certain symbolic outputs from certain symbolic inputs. That's what algorithms do. (An intuitive example is the rule we all learned in school for extracting the roots of quadratic questions: "minus *b* plus or minus the square root of ….")

Algorithms are like recipes: apply them to the symbolic ingredients and you can explain how to bake a symbolic cake. Computation is very powerful; just about everything in the universe can be encoded symbolically and explained computationally, including cardiac function. The right algorithm can pump symbolic blood. And you can show that the algorithm really works by applying it to build a synthetic heart that really passes the cardiac Turing Test (TT) and pumps blood. But to do that, you have to "interpret" the symbolic code and implement it in material form, just as a formal recipe for a cake needs to be implemented in material form, using the real ingredients referred to by the symbols, to generate a real cake.

So, despite its enormous power, computation cannot be all there is to cognition. Searle (1980) showed, famously (in this journal), that a computer is not cognizing even if it can pass the TT because Searle too could pass the Chinese TT by executing the symbolic code without understanding a word of Chinese. Why can't he understand? Because there is no connection between the symbols in the code and the objects in the world that they are interpretable